UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model

Year: 2023 | *Citations: 122* | *Authors: Jiabo Ye, Anwen Hu, Haiyang Xu*

Abstract

Text is ubiquitous in our visual world, conveying crucial information, such as in documents, websites, and everyday photographs. In this work, we propose UReader, a first exploration of universal OCR-free visually-situated language understanding based on the Multimodal Large Language Model (MLLM). By leveraging the shallow text recognition ability of the MLLM, we only finetuned 1.2% parameters and the training cost is much lower than previous work following domain-specific pretraining and finetuning paradigms. Concretely, UReader is jointly finetuned on a wide range of Visually-situated Language Understanding tasks via a unified instruction format. To enhance the visual text and semantic understanding, we further apply two auxiliary tasks with the same format, namely text reading and key points generation tasks. We design a shape-adaptive cropping module before the encoder-decoder architecture of MLLM to leverage the frozen low-resolution vision encoder for processing high-resolution images. Without downstream finetuning, our single model achieves state-of-the-art ocr-free performance in 8 out of 10 visually-situated language understanding tasks, across 5 domains: documents, tables, charts, natural images, and webpage screenshots. Codes and instruction-tuning datasets will be released.